IN THE UNITED STATES DISTRICT COURT
FOR THE SOUTHERN DISTRICT OF NEW YORK

HACHETTE BOOK GROUP, INC., HARPERCOLLINS PUBLISHERS LLC, JOHN WILEY & SONS, INC., and PENGUIN RANDOM HOUSE LLC Plaintiffs, v. INTERNET ARCHIVE and DOES 1 through 5, inclusive Defendants.	Case No. 1:20-CV-04160-JGK

DECLARATION OF IAN FOSTER

Pursuant to 28 U.S.C. § 1746, I, **IAN FOSTER**, declare as follows:

1. I am the Arthur Holly Compton Distinguished Service Professor of Computer Science at the University of Chicago and a Distinguished Fellow, Senior Scientist, and Division Director at Argonne National Laboratory. I submit this declaration in connection with the motion for summary judgment filed by plaintiffs, Hachette Book Group, Inc., HarperCollins Publishers LLC, John Wiley & Sons, Inc., and Penguin Random House LLC (collectively, "Plaintiffs"), in their lawsuit against Defendant Internet Archive ("IA"). My statements set forth below are based upon my specialized knowledge, education, and experience, as applied to the facts and circumstances of this case. If called upon, I would and could competently testify as to the matters contained herein.

I. My Assignment

2. Plaintiffs in this action have asked me to conduct a technical analysis concerning IA's digitization of print books and use of the resulting digital books. My assignment includes providing an overview of that analysis and considering the following five questions:

- Where does copying and distribution of books occur within IA's processes?
- How has IA copied and disseminated the books listed in the Complaint?
- How has IA copied and used books from Plaintiffs' respective in-print catalogs?
- How has IA's distribution of digitized books evolved over time?
- Are IA's methodologies for digitizing print books and then lending the resulting digital books well-developed and reliable from a technological perspective?

3. In my work in this matter, I have reviewed and utilized a variety of materials, as cited specifically in my expert report and supplemental expert report. This includes, but is not limited to, (a) various publicly accessible websites, including IA's sites at www.archive.org and www.openlibrary.org; (b) numerous documents that IA produced in discovery, including source code, technical documents, data sets, and digital versions of books; and (c) deposition testimony.

4. In discussing my analysis in this declaration, I do so in a manner geared toward a general audience. My report and supplemental report describe my analysis in more technical terms, with detailed citations, but I assume the higher level of discussion herein is more assistive to the Court.

II. My Background

5. I am the Arthur Holly Compton Distinguished Service Professor of Computer Science at the University of Chicago and a Distinguished Fellow, Senior Scientist, and Division Director at Argonne National Laboratory. In these capacities, I teach students, conduct research with collaborators at these institutions and others, and direct a division at Argonne focused on data science and machine learning. I often take the role of Principal Investigator, securing funding grants from federal science & technology agencies and leading research programs. Prior to joining the University of Chicago and Argonne roughly three decades ago, I obtained a Bachelor of Science (B.Sc. Hons I) degree in Computer Science from the University of Canterbury

(Christchurch, New Zealand), and a Doctor of Philosophy (Ph.D.) degree in Computer Science and Diploma of Imperial College from Imperial College (London, United Kingdom).

6. My work as a computer scientist has been at the intersection of computing and the sciences. My work has produced both practical technologies that have seen wide adoption and concepts and methods that have proven influential in research and education. This includes founding the Distributed Systems Laboratory at Argonne and the University of Chicago, where I established the multi-institutional Globus project (www.globus.org). Globus has developed core technologies for management and transfer of distributed research data at the largest scales and is an active non-profit service for use by the research community at institutions worldwide. The resulting technologies formed the basis for many national and international "Grid" projects—now considered as foundational aspects of "cloud computing" for science and engineering—funded by the Department of Energy (DOE), NASA, the National Science Foundation (NSF), the European Union, and the UK eScience Program. I have published eight books, over 300 journal articles, and over 300 articles in conference and workshop proceedings. My publications have been cited more than 130,000 times. My full curriculum vitae is attached as **Appendix A** to this declaration.

III. IA's Book System in Context

7. IA offers several different online services. My analysis in this matter concerns IA's services pertaining to books and the technological implementation of these services, which I refer to as the "Book System."

8. The Book System comprises both frontend elements (web pages) via which users interact with the system, and backend components for, among other things, digitizing print books, maintaining a catalog of digitized books, and maintaining the system. Though the technical analysis may seem complicated, the process of obtaining digital books from IA is seamless to the

user. Any Internet-connected user can quickly sign up for an IA account and within moments obtain digital books in whole for reading.

9. The scale of IA's Book System is vast: IA has made eBooks that are wholesale copies of millions of physical books of all types, including fiction and non-fiction. IA disseminates the digital books to registered users of its website. Anyone, anywhere in the world, can sign up for a free account with IA and receive free access to complete copies of books that are currently commercially available.

10. IA's Book System contains both digital copies of books uploaded by users and digital books generated by IA by scanning physical books. My analysis in this matter concerns the latter.

11. IA's Book System contains both digital copies of books that are in the public domain and digitized modern books that are still under copyright protection. My analysis in this matter concerns the latter content, which appears within an area of IA's website that IA describes as its "inlibrary" (or "Books to Borrow") collection.

12. A premise to how IA operates its Book System is that IA generally restricts the number of scanned copies of a particular book accessible from its website at any one time to not more than the number of physical copies of that book that IA and its partner libraries believe they own. IA calls this a 1:1 "owned to loan ratio." Below, I explore IA's technical approach to that concept, including IA's use of what it calls its "Open Libraries" program with partner libraries, under which IA lends multiple copies of a single scan of a particular book at a time based on the number of print copies that the partner libraries reported they have in their possession. I also explore the history of IA's copying and distribution of Plaintiffs' literary works generally and works-in-suit ("WIS") specifically.

13. At the outset, it is important to recognize that IA has made changes to its Book System over time, including during this litigation. The size of the inlibrary collection has grown significantly within the past few years. IA's approach to providing users with access to digital books has also changed and could change further at any time. For instance, shortly after the litigation was filed in June 2020, IA introduced the concept of 1-hour loans rather than just 14-day loans. Prior to June 2020, each loan of a digitized book in the inlibrary collection was for a 14-day period only. By way of another example, for a roughly three-month period that commenced on or about mid-March 2020, IA announced what it termed a "National Emergency Library" whereby IA altered how its website operates, so that IA could distribute a virtually limitless number of copies of any book in its inlibrary collection, no matter the number of print copies on hand.

IV. Summary of Certain Opinions

14. As mentioned, I was asked to conduct a technical analysis concerning IA's digitization of print books and use of the resulting digital books. Part of that assignment included analyzing five specific questions, which I address below along with a summary of my answers to these questions. A more detailed discussion is contained in the sections that follow afterwards.

Q1: Where does copying and distribution of books occur within IA's processes?

15. IA makes a copy of print books when it scans those books to create a digital version of a physical book. The initial copy consists of JPEG images on IA's Scribe machines. A subsequent copy is made when uploaded to IA's servers. IA then makes over 10 different copies of that scan in various formats. *See* ¶¶ 50-66. IA distributes yet additional copies of the digitized book to users, either for display in their online browser (in IA's BookReader) or in the form of a download for offline reading. Copies are also made as users employ the listen feature of IA's BookReader, in order to have the book read to them. IA copies and disseminates the digital books in their entirety

(or, in the case of BookReader access, in response to user navigation actions), so that registered users of IA's Site can read the complete book. *See* ¶¶ 20-37.

Q2: What is the history of IA copying and distributing the books listed in the Complaint?

16. For each WIS (i.e., the books listed in Exhibit A to the Complaint), IA has made one or more initial scanned copies; uploaded each of those copies to its servers; from those copies then made 10 or more different digital copies in various formats; and then provided the digital books to users. In almost all instances, IA provided the book for digitization and either IA or the Kahle/Austin Foundation sponsored the digitization. The loan information that IA provided depicts IA lending copies of each WIS to users, with some of those works having been the subject of thousands of loans, and a total of 46,307 loans of a WIS. The information that IA provided also shows many instances of IA providing multiple users at a time with access to copies of a single scan of a particular WIS (both before and during the period IA calls the "National Emergency Library"). The data that IA produced is for a limited snapshot in time (i.e., roughly March 2017 until early September 2020); as a result, for some scans, the data IA provided is not a full accounting of all loans and additionally may not reflect the actual maximum number of users who had access to copies at a time ("actual concurrent loans"). *See* ¶¶ 96-107 and 111.

Q3: How has IA copied and used books from Plaintiffs' respective in-print catalogs?

17. As I describe below, my analysis reflects that the WIS are essentially the "tip of the iceberg" for each Plaintiff. IA has digitized, and is distributing to users, substantial portions of each Plaintiff's respective in-print (i.e., commercially available) catalogue. In addition to the WIS, IA has scanned and has in its Books to Borrow collection 33,003 other books from Plaintiffs' in-print catalogues. Those are books for which Plaintiffs offer both a physical book and an eBook or audio book. *See* ¶¶ 112-118. Further, for certain WIS and other of Plaintiffs' works, IA has

received notices of alleged infringement but, by not removing or disabling access to the item identified in the notice, its actions have fallen short of its technical capabilities. *See* ¶¶ 119-135.

Q4: How has Internet Archive's distribution of digitized books evolved over time?

18. As I explain below, my analysis is that IA's scanning of physical books and online dissemination of the resulting digital books has expanded significantly, especially since 2018. The number of scans in IA's inlibrary collection has increased from (a) 648,117 scans on April 1, 2018; (b) 965,499 scans on April 1, 2019; (c) 1,476,344 scans on May 26, 2020; and (d) 3,211,204 scans on February 19, 2022. *See* ¶ 22. The data further reflects heavy scanning of Plaintiffs' works in 2020 and 2021, as compared to prior years. *See* ¶ 118. In addition, I note that, for many of these digital books, IA provides multiple users with copies of the book at the same time. The impact of IA's willingness to make simultaneous loans of a single scan can be seen in my exhibits that lay out, for each WIS, the maximum eligible concurrent lending limit and the maximum actual concurrent loans made, both of which often have a value much greater than one. Indeed, as demonstrated by Exhibit 5 and Exhibit 5A, there are many WIS for which IA's records show more than 10 eligible concurrent loans as a result of the Open Libraries overlap analysis. For example, for *Ship Breaker* by Paolo Bacigalupi (WIS #4), Exhibit 5A indicates 12 eligible concurrent loans that IA will lend out at any given time based on its one scan. Another example is *Redeployment* by Phil Klay (WIS #55); Exhibit 5A indicates IA permits 17 concurrent loans from one scan. IA compounds this effect by maintaining multiple distinct scans for a single work. Consequently, Exhibit 5A shows that IA allows 40 concurrent loans of *The Lion, the Witch, and the Wardrobe* by C.S. Lewis (WIS #70) and 23 concurrent loans of Bill Bryson's *A Short History of Nearly Everything* (WIS #9).

Q5: Are IA's methodologies for digitizing print books and then providing the digital copies for lending well-developed and reliable from a technological perspective?

19. IA digitizes books systematically and at enormous scale. However, I have observed problems that occur thereafter from a technological perspective. I discuss this topic in several sections below. I list some of my analysis and opinions here as follows:

- IA does not have a rigorous system for ensuring the reliability of metadata. That, in turn, leads to errors in IA's understanding as to the contents of its digitized scans, how many loans it has made, and how many copies of a particular book are eligible (under IA's policies) for user access at a time. *See* ¶¶ 86-95.

- IA's methodologies are inconsistently applied. For instance, while IA has a policy regarding what can be put into the inlibrary collection, and has a script that applies that policy, IA sometimes fails to adhere to it. *See* ¶¶ 67-69.

- IA's approach to overlap analyses with partner libraries (a) treats all copies of a book as interchangeable items, without regard to differences in their respective physical condition; (b) inflates the likelihood of finding a match by using excess ISBNs that do not correlate with IA's copy; (c) does not verify that the library actually possesses what its records depict; and (d) does not stay in synch with events such as the library discarding, losing, or loaning its book, or a local patron reading it in the physical library. *See* ¶¶ 44; 75-85.

- When IA chooses to do so, it provides many users at a time with copies of IA's scan of a particular book. *See* ¶¶ 42-43; 48.

V. Accessing Books in the Book System

20. A user may approach the Book System in a variety of ways. Two significant ways are via the main IA website (https://archive.org) (the "IA Site") and IA's Open Library site (https://openlibrary.org) (the "OL Site"). Users may also arrive at the Book System in other ways,

such as through an internet search; a library's website; the website of IA's affiliate, Better World Books; or through other websites that link to the Book System.

A. Main Internet Archive site

21. The IA Site's home page, shown in *Figure 1*, provides access to a variety of IA's services. The header includes a "BOOKS" tab that, when clicked, prominently shows icons for "Books to Borrow" and "Open Library", as shown in *Figure 2*. The first icon leads to a page for the "Books to Borrow" collection of electronic books on the main IA Site, as shown in *Figure 3*. The second leads to the distinct OL Site.

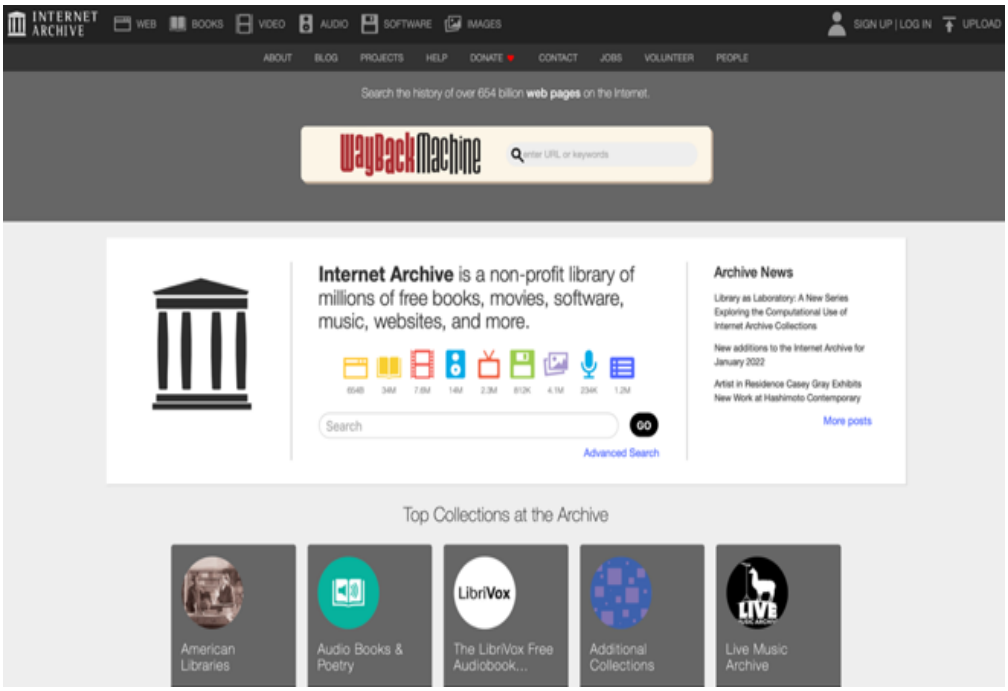


Figure 1: IA Site home page, at https://archive.org. Note the "Books" item in the upper left menu bar.

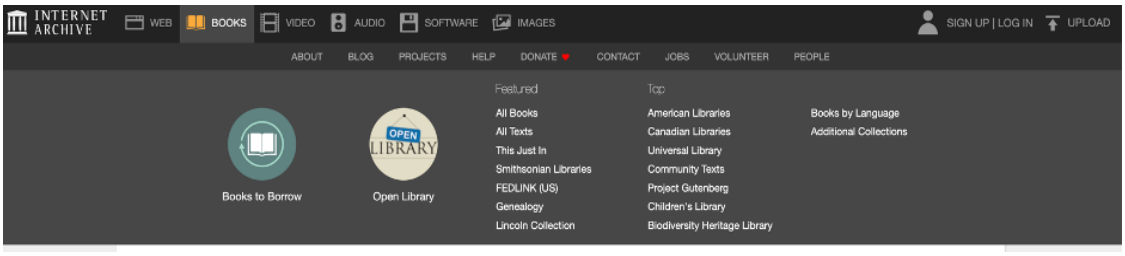


Figure 2: Expanding the "Books" menu item in Figure 1 reveals two IA book services: "Books to Borrow" and "Open Library."

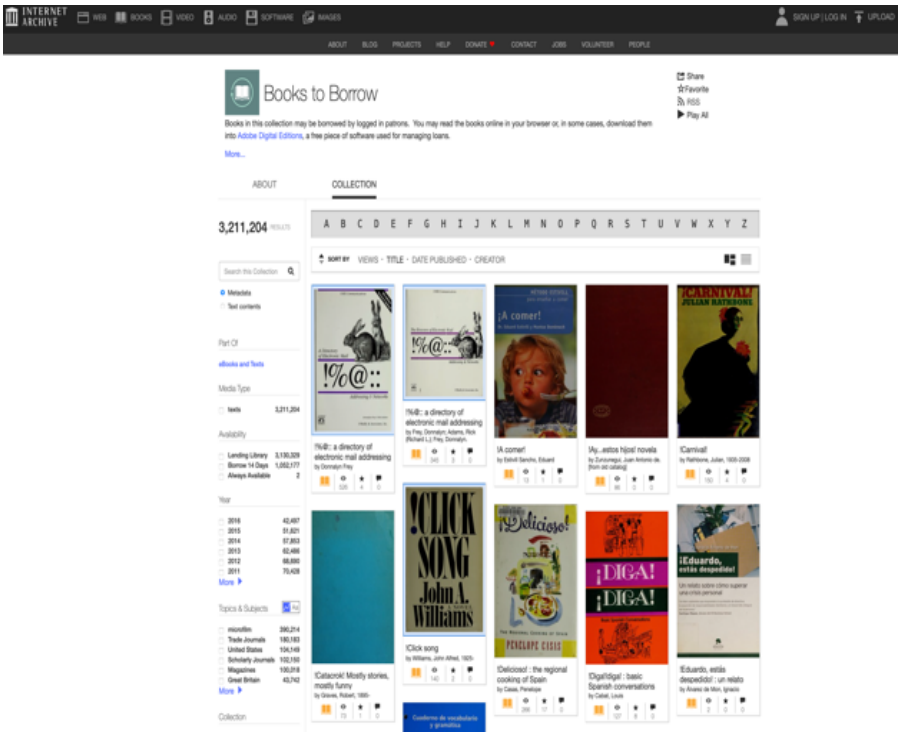


Figure 3: The Books to Borrow main page, at https://archive.org/details/inlibrary, as of February 21, 2022.

22. IA organizes its various content into "collections." The "Books to Borrow" page provides access to digitized books within the "inlibrary" collection, which the page shown in *Figure 3* reported as holding 3,211,204 items as of Feb 21, 2022. The collection is named "inlibrary" because its earlier version was a literal "In-Library Lending" program, in which scans were provided to users of computers physically located within libraries partnering with Internet Archive. In 2011, the details page for the inlibrary collection described it as "Books available to patrons physically present in designated libraries." IA's Wayback Machine online repository tracks the appearance of the inlibrary collection URL at different points in time. While the inlibrary collection currently contains over 3 million scanned copies of books: on May 26, 2020, in the week before the Plaintiffs filed this lawsuit, the collection contained 1,476,344 scans; on April 1, 2019, the collection contained 965,499 scans; and on April 1, 2018, the collection contained 648,117 scans.

23. Clicking on the "About" icon on the "Books to Borrow" page shown in *Figure 3* toggles to text explaining that "Books in this collection may be borrowed by logged in patrons. You may read the books online in your browser or, in some cases, download them into Adobe Digital Editions, a free piece of software used for managing loans." The description at the "About" icon on the "Books to Borrow" page, at https://archive.org/details/inlibrary?tab=about, indicates:

> Books in this collection may be borrowed by logged in patrons. You may read the books online in your browser or, in some cases, download them into Adobe Digital Editions, a free piece of software used for managing loans.
>
> Please note that works in this collection are protected by copyright law (Title 17 U.S. Code) and copying, redistribution or sale, whether or not for profit, by the recipient is not permitted unless authorized by the rightsholder or by law.
>
> See FAQs about borrowing books.
>
> Libraries can participate in our Open Libraries program and lend these digital titles to their patrons by filling out our online form.

24. As I explain herein, until June 2020, the IA Site's functionality enabled downloads of any borrowed book into Adobe Digital Editions. During this case, in June 2020, IA changed its system to introduce 1-hour loans for some books, with copies borrowed on 1-hour loans available only for online viewing in your browser. In any event, users can use a search feature on the "Books to Borrow" page shown in *Figure 3* to locate titles within the collection. Apart from that search interface, IA groups books into various "Topics and Subjects" for users to find a book.

25. For example, a search within Books to Borrow for "*Triumph of the City Glaeser*" returns a single result, represented by a thumbnail image of the book's cover, its title (*Triumph of the City*), and the author's name (Edward Glaeser). Clicking on the thumbnail image or associated text for this title takes the user to the web page at address "https://archive.org/details/isbn_9780143120544," an address that comprises three components: the domain at which the page is hosted ("archive.org"); the word "details"; and a unique identifier assigned by Internet Archive to this scan, "isbn_9780143120544."

26. Visiting the web page at https://archive.org/details/isbn_9780143120544 displays a reader screen in which the user can read the book. This "BookReader application" allows the user to use the arrow keys on their keyboard, or to click on navigation buttons in the reader, to move to different pages in the book.

27. IA requires users to be logged in to an IA account to access many of its functions, including full access to free copies of complete digital books in the inlibrary collection. For example, if I navigate forward through *Triumph of the City* without logging in, I am soon told to "use your free account to borrow this book and gain access to all pages". If I log in, and then borrow the digital book for either a 1-hour loan or 14-day loan, I can access the remainder of the book.

28. The "BookReader application" allows the user to use the arrow keys on their keyboard, or to click on navigation buttons in the reader, to move to different pages in the borrowed book. The BookReader app on the IA Site also has a "Read Aloud" feature, which a logged in user can activate for a borrowed book by clicking on a headphones icon in the BookReader. The feature converts the text to audio and plays it aloud.

29. Scrolling down the screen at https://archive.org/details/isbn_9780143120544, displays information about the book (*Figure 4*), including text on "DOWNLOAD OPTIONS". Users can download the borrowed book as an EPUB file or a PDF file. A "14 day loan [is] required to access EPUB and PDF files." There is also a special format, the "ENCRYPTED DAISY" file, for print-disabled users with special accounts.

Triumph of the city : how our greatest invention makes us richer, smarter, greener, healthier, and happier

by Glaeser, Edward L. (Edward Ludwig), 1967-

☆ Favorite | Share | Flag

Publication date	2011
Topics	Urbanization, Cities and towns, Urban economics, Sociology, Urban, Urbanisation, Villes, Economie urbaine, Sociologie urbaine, Cities and towns, Sociology, Urban, Urban economics, Urbanization, Stedelijke economie, Forensisme, Stadscultuur, Stadtentwicklung, Stadtsoziologie, Stadtökonomie, Verstädterung, Wachstum, Stadtentwicklung, Stadtsoziologie, Stadtökonomie, Verstädterung, Wachstum, Urbanization, Cities and towns, Cities and towns, Urban sociology
Publisher	New York : Penguin Press
Collection	inlibrary; printdisabled; internetarchivebooks; china
Digitizing sponsor	Kahle/Austin Foundation
Contributor	Internet Archive
Language	English

Includes bibliographical references (pages 307-323) and index

A pioneering urban economist offers fascinating, even inspiring proof that the city is humanity's greatest invention and our best hope for the future

Our urban species -- What do they make in Bangalore? -- Why do cities decline? -- What's good about slums? -- How were the tenements tamed? -- Is London a luxury resort? -- What's so great about skyscrapers? -- Why has sprawl spread? -- Is there anything greener than blacktop? -- How do cities succeed? -- Flat world, tall city

Our urban species -- What do they make in Bangalore?: Ports of intellectual entry: Athens ; Baghdad's house of wisdom ; Learning in Nagasaki ; How Bangalore became a boom town ; Education and urban success ; The rise of Silicon Valley ; The cities of tomorrow -- Why do cities decline?: How the rust belt rose ; Detroit before cars ; Henry Ford and industrial Detroit ; Why riot? ; Urban reinvention: New York since 1970 ; The righteous rage of Coleman Young ; the Curley effect ; The edifice complex ; Remaining in the rust belt ; Shrinking to greatness -- What's good about slums?: Rio's favelas ; Moving on up ; Richard Wright's urban exodus ; Rise and fall of the American ghetto ; The inner city ; How policy magnifies poverty -- How were the tenements tamed?: The plight of Kinshasa ; Healing sick cities ; Street cleaning and corruption ; More roads, less traffic? ; Making cities safer ; Health benefits -- Is London a luxury resort?: Scale economies and the Globe Theatre ; The division of labor and lamb vindaloo ; Shoes and the city ; London as marriage market ; When are high wages bad? -- What's so great about skyscrapers?: Inventing the skyscraper ; The soaring ambition of A.E. Lefcourt ; Regulating New York ; Fear of heights ; The perils of preservation ; Rethinking Paris ; Mismanagement in Mumbai ; Three simple rules -- Why has sprawl spread?: Sprawl before cars ; Arthur Levitt and mass-produced housing ; Rebuilding America around the car ; Welcome to The Woodlands ; Accounting for tastes: why a million people moved to Houston ; Why is housing so cheap in the sunbelt? ; What's wrong with sprawl? -- Is there anything greener than blacktop?: The dream of garden living ; Dirty footprints: comparing carbon emissions ; The unintended consequences of environmentalism ; Two green visions: the prince and the mayor ; The biggest battle: greening India and China ; Seeking smarter environmentalism -- How do cities succeed?: The Imperial city: Tokyo ; The well-managed city: Singapore and Gaborone ; The smart city: Boston, Minneapolis, and Milan ; The consumer city: Vancouver ; The growing city: Chicago and Atlanta ; Too much of a good thing in Dubai -- Flat world, tall city: Give cities a level playing field ; Urbanization through globalization ; Lend a hand to human capital ; Help poor people, not poor places ; The challenge of urban poverty ; The rise of the consumer city ; The curse of NIMBYism ; The bias toward sprawl ; Green cities ; Gifts of the city

576 Previews
6 Favorites

PURCHASE OPTIONS
Better World Books

DOWNLOAD OPTIONS
ENCRYPTED DAISY 1 file
For print-disabled users

14 day loan required to access EPUB and PDF files.

IN COLLECTIONS
Books to Borrow
Books for People with Print Disabilities
Internet Archive Books
Scanned in China

Uploaded by
Tracey Gutierres
on May 21, 2015

Figure 4: Various descriptive information about the book.

30. With a 14-day loan, I am presented with additional "DOWNLOAD OPTIONS" (*Figure 5*), namely "ENCRYPTED ADOBE EPUB" and "ENCRYPTED ADOBE PDF" (with "High Quality Page Images"). Clicking on the latter results in a file download to my computer, which I can open in the "Adobe Digital Editions 4.5" application. The scanned book is then in my Adobe Digital Editions "Library," from where I can open and read it on my computer and a variety of other devices (e.g., an iPad).

Figure 5: Once I have borrowed the book "Triumph of the City" for 14 days, the "DOWNLOAD OPTIONS" (center right) changes to include additional options.

31. Multiple users can obtain and read copies of the same IA scan of a book at the same time. For example, I verified that I could log in to the IA Site with two other accounts and obtain a second and third 14-day loan for *Triumph of the City*, for a total of three 14-day loans active at the same time. (I did not try more than three accounts total.)

32. Each digital book that users access via the IA Book System is actually a copy of a scan of a physical work. Each scan is identified by a unique identifier generated by IA. For example, the web page shown in *Figure 4* has internet address, https://archive.org/details/isbn_9780143120544. Here, the prefix "isbn_9780143120544" corresponds to the IA identifier for this specific scan of the book, *Triumph of the City*.

33. While a search for the title *Triumph of the City* reveals a single scan, other titles may have several. For example, a search for "*Rogue Lawyer* Grisham" reveals eight scans: see *Figure 6*.

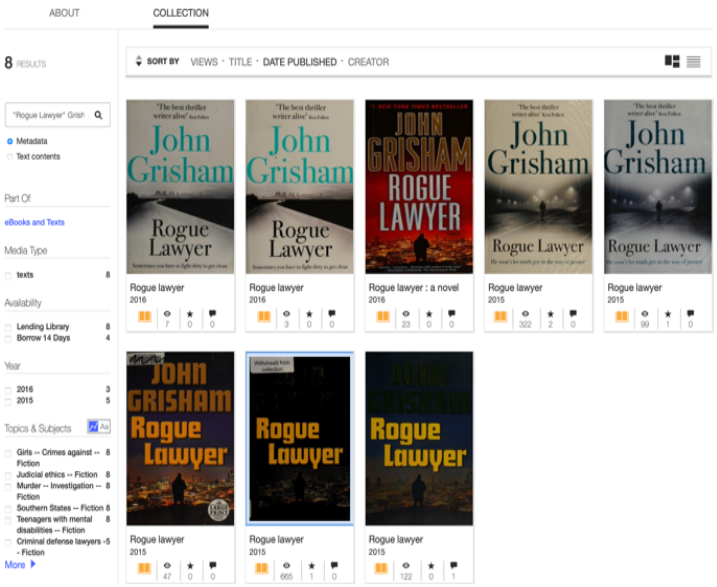
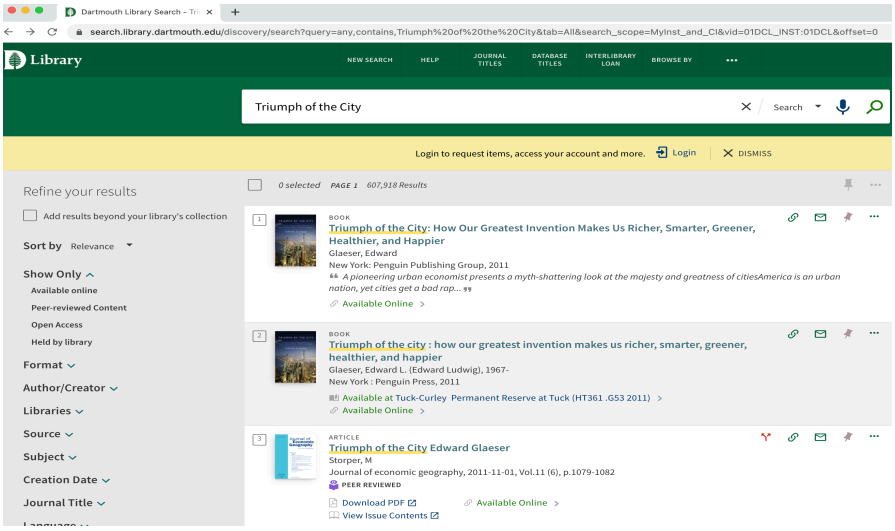


Figure 6: A search for '"Rogue Lawyer" Grisham' indicates eight matching scans.

B. Access from Library Sites

34. Certain libraries integrate links to the IA Site into their website search systems, so that users searching for certain titles are directed to the IA Site. In such instance, a person searching their library's website search system may be referred to the IA Site for a digital copy of a book. Dartmouth Library is an example. *Figure 7* shows me accessing, from the Dartmouth Library search page, the familiar *Triumph of the City*.



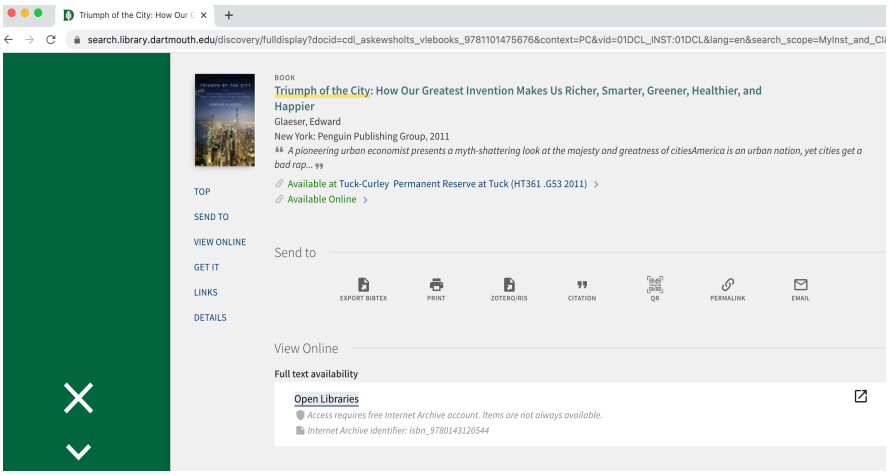


Figure 7: From top to bottom: I search for Triumph of the City book on Dartmouth Library web site; I am told that the full text is located in Open Libraries.

C. Internet Search

35. Internet search engines index IA Site pages. Thus, a person searching a search engine may be referred to the IA Site for a digital copy of a book.

D. Access from Better World Books

36. Links to IA's digital copies of books are also integrated into an online books platform called Better World Books ("BWB"), which is owned by an affiliate (Better World Libraries) of Internet Archive. BWB offers users seeking to purchase a new or used book the option to download or display it instead via IA for free. The webpage where BWB offers books (new and used) for sale also includes a "Borrow" button to obtain a "Digital edition from Internet Archive."

E. The Open Library Site

37. The other prominent icon on the IA Site, as seen in *Figure 2*, takes the user to https://openlibrary.org, an affiliated website, the "OL Site", also operated by IA. The OL Site supplies information on individual books. Each page is comparable to a card in a library card catalog. The OL Site maintains web pages, each of which—in a manner akin to a card in a library card catalog—provides information about a published work and associated editions of that work.

Many of these pages include links to copies ("scans") of specific editions that are available electronically from the IA Site. Links to preview, borrow, or listen to a book lead to the IA Site's BookReader application. Borrowing requires an account with the IA Site (or OL Site).

VI. Technical implementation supporting access to books

38. I now discuss some aspects of the technical implementation of the Book System and how it supports access to, and dissemination of, digital books. As discussed above, these digital books are copies ("scans") of print books.

39. IA assigns each distinct scan of a book a unique identifier, such as the identifier "isbn_9780143120544" noted above for *Triumph of the City*. Internally, IA maintains a mapping from each such identifier to the associated digital copy, plus various information about the book that the scan represents, the scanning process, and so forth: what is referred to collectively as "metadata." IA allows for multiple digital items to be organized into groupings called "collections." A collection has a set of digital items, an administrator, and permissions settings controlling who can access its items.

40. Two collections of relevance to my analysis are the "inlibrary" and "printdisabled" collections. Items in both these collections are what IA calls "access restricted." For the "inlibrary" collection, a user must log in and then "borrow" a scan in order to view or listen to a copy of that scan. Once a user has logged in, whether they can borrow a specific "inlibrary" scan depends primarily on the number of copies that IA has recorded as being available for that scan. That availability is a function of the existence of a physical copy under IA's control or (as explained below) due to a library reporting it also has a copy that can contribute to the lending count, and the number of other concurrent loans for that scan. The loan is allowed if the number of currently

active loans for the requested scan is less than the number of copies of that digital book that IA has recorded as being available for lending.

A. Control Over Current Access to a Scan

41. In general, any scan in the "inlibrary" collection is available for access by a registered IA user: that is, by any individual who has created an account on the IA Site (or the OL Site). But whether a user can access a particular scan at a particular point in time is determined by the "maximum eligible concurrent loan limit" that IA has defined for that scan, and the number of accesses to that scan that are currently in progress: in IA's terminology, the number of active 1-hour and 14-day loans for the scan. In general, according to the architecture of IA's Book System, except other than when IA chooses to proceed differently, or has incorrect metadata, the number of concurrent accesses should not exceed the maximum eligible concurrent loan limit.

42. IA sets the concurrent loan limit for a particular scan to 1 plus a "number contributed by partners." The overall idea is that the "1" corresponds to the physical copy of a particular book that IA has recorded that it possesses, and has scanned, and the "number contributed by partners" corresponds to physical copies of the same edition of the same book thought to be located on the partners' bookshelves or stacks.

43. The "copies contributed by partners" is thus a vital element of IA's approach to disseminating digital copies of scans in its inlibrary collection, in that IA uses this information to justify increasing its maximum eligible concurrent loan limit beyond the number of copies of a book that IA believes it owns. IA obtains this information from partners via a process called "overlap analysis," which I discuss in more detail below.

44. IA uses an overlap analysis to compare a list of items that metadata indicate are present in IA's "inlibrary" collection with a list of items that metadata indicate are present in a collection of

books that a partner library intends to "make available" to IA, to determine which items are found in the metadata associated with both collections. This comparison is performed on the basis of metadata, not the physical items; thus, it is only as accurate as the metadata that IA maintains itself, or that is provided to it by third parties. Furthermore, the analysis assumes that works with matching metadata are interchangeable, regardless of their physical condition. To determine which scans in the inlibrary collection match with works found in a particular partner's collection, IA obtains from the partner a list of items thought to be in the partner's collection, and then performs overlap analysis.

45. As noted earlier, the Book System on the IA Site supports both 1-hour loans and 14-day loans. (Initially, only 14-day loans were supported; 1-hour loans—also referred to as "session loans"—were introduced in June 2020.) Fourteen-day loans are offered if IA has set the book's maximum eligible concurrent loan limit to be more than one. If the maximum eligible concurrent loan limit is 1 only, the Book System permits only 1-hour loans.

46. A consequence of the "maximum eligible concurrent loan limit" is that a user request to access a copy of a scan may be declined. IA's Book System then allows a user to be placed on a "waitlist," to be notified when the scan is available.

47. I reviewed the source code used to implement the loan logic in the IA Site. To be lendable, a scan must be in the "inlibrary" collection. The code then determines how many copies can be loaned by counting the number of ISBN matches in different collections. Specifically, it counts one for each collection in which the scan is located.

B. National Emergency Library

48. IA disabled its restrictions on the number of concurrent loans limits for books in the "inlibrary" collection during a period in 2020, a program that IA referred to as the "National

Emergency Library" ("NEL"). To be more precise, IA's implementation of this mechanism involved not disabling the restrictions but rather artificially raising the number of "contributed items" to be considered when determining whether to grant a request to access a scan to a high number—but the effect was in practice the same. IA personnel have also described the National Emergency Library as "suspending waitlists."

VII. Backend Elements of the Book System

49. There are several backend elements that support the ability for users to access copies of books. IA has changed the system over time and continues to make changes; it can change the rules at any time, such as seen with the introduction and removal of the NEL; the addition of 1-hour loans; and the expansion of digitize-and-lend to include the Open Libraries initiative. Indeed, the name "inlibrary" derives from IA's original rollout of providing digital copies of books under a technological requirement that the user must be "in the library" rather than the IA Site providing copies to anyone, anywhere.

A. Adding Books

50. IA acquires physical books and digitizes them by taking digital photographs of their covers and all of their pages. It uses custom-built "Scribe" workstations for this purpose. The scanning process generates images of each page of the book in JPEG format, from which IA generates copies of the scanned book in various formats, including EPUB and PDF. This process is also called "scanning" the book, and the resulting digital version is referred to as a "scan."

51. IA has Scribe workstations at locations throughout the United States and other countries, often embedded within libraries. IA also arranges for companies in the Philippines and Hong Kong (or mainland China) to digitize books on IA's behalf.

52. The digitization process is non-destructive. After photography, IA ships the physical

book to a storage facility and places it in a container with a "Box ID" number.

53. IA assigns each scan a unique identifier, which is a short alphanumeric string.

54. For each scan, IA maintains metadata records. These records can be said to be "attached" to the scan within IA's system; that is, the metadata are associated with the scan identifier. These metadata include a number of operational facts, such as the model of camera used and the date of the scan, as well as the Box ID of the scanned book.

55. IA also attaches to each scan metadata regarding the content of the book, such as its title, author, date, and publisher.

56. Photography of the book using a Scribe machine results in a set of digital image files in the JP2 (JPEG 2000) format. IA uploads them to its servers in California. On IA servers, the filenames carry the suffix "orig" to indicate these are the original images. This is a **first** digital copy of the book.

57. IA then processes the original image files, carrying out a number of steps to yield files for users of the Book System.

58. First, IA rotates, crops, and performs other image adjustments as necessary to the original JP2 files. This yields a **second** digital copy of the book, in the form of a second set of JP2 files. I will refer to this as the set of adjusted images.

59. IA then proceeds to use the adjusted images to create several more copies. It also delivers the adjusted images to users one image at a time as they page through books using BookReader on the IA Site.

60. IA combines the adjusted images into a single PDF file, which is a **third** copy of the book.

61. IA generates an encrypted version of the PDF. This is the file IA distributes for reading with the Adobe Digital Editions application. This is a **fourth** copy of the book.

62. IA uses optical character recognition (OCR) software to extract text from the images of the pages. This software automatically identifies characters and their positions on each page and generates a purely textual, as opposed to image-based, copy of the book's content. The text is stored in simple, readable form in a file with filename extension "_djvu.txt". This is a **fifth** copy of the book. The characters and data on their positions on the pages also are stored in two formats designed for use by software; these files have filename extensions "_djvu.xml" and "_abbyy." These are **sixth** and **seventh** copies of the book.

63. In addition to the PDF, IA generates another version for download in a format known as EPUB. This version is a composite of text from the OCR process and images (such as figures and illustrations). IA also creates an encrypted version of the EPUB file that can be read only with Adobe Digital Editions application. These are **eighth** and **ninth** copes of the book.

64. IA also creates a copy in the DAISY format, which as discussed earlier, is a version for use by print-disabled users with special accounts. IA also generates an encrypted version of the DAISY copy. These are **tenth** and **eleventh** copies of the book.

65. In some cases, IA inserts an Internet Archive "bookplate" page into the beginning pages of the scan indicating the book was digitized by IA. Other than that, IA scans the complete book and without adding any content to the book.

66. IA refers to the technical creation of these multiple copies as the "derive" process, because the various copies are derived from the original photos of the book taken with the Scribe machine. More generally, IA uses the term "republishing" to refer to the full process, including manual quality assurance steps, of moving from the original images to the copies provided to users.

B. Making Books Lendable

67. After IA adds a scan to the Book System, it may decide to make the scan available for

lending, which means it is available for reading either on the IA Site or on the user's device for a limited time, either one hour or fourteen days. To make a scan lendable, IA assigns the scan to the "inlibrary" collection, also known as the "Books to Borrow" collection. IA accomplishes this assignment by updating the "collection" metadata item attached to the scan. The "collection" item consists of a list of one more IA collections to which the scan belongs.

68. When a new book is scanned, it is typically first placed in an access-restricted collection called the "print-disabled" collection. From there, it may be moved to the "inlibrary" collection.

69. IA's protocols indicate that books published within the past five years should not be made available for lending. However, I have observed that IA sometimes deviates from the date criteria set forth in the document described in the prior paragraph. Here are two examples involving a WIS where IA scanned a book and put it within its inlibrary collection much sooner:

- The identifier allpresidentswom0000levi is a copy of the book *All the President's Women: Donald Trump and the Making of a Predator* by Barry Levine and Monique El-Faizy (WIS #69). The book was published in 2019, and the IA metadata also reflects that the book was published in 2019, but IA added it to the inlibrary collection on October 28, 2019.

- The identifier manwhosolvedmark0000zuck is a copy of the book *The Man Who Solved the Market: How Jim Simons Launched the Quant Revolution* by Gregory Zuckerman (WIS #127). The book was published in 2019, and the IA metadata also reflects that the book was published in 2019, but IA added it to the inlibrary collection on December 9, 2019.

C. Increasing concurrent lending limits: Open Libraries project

70. As I described in the previous section, when IA scans a physical book, places it in storage, and then adds the scan to the inlibrary collection, the scan becomes lendable to users. When a user takes such a loan, the scan becomes unavailable to other users until the book is returned or the loan expires. At any given time, except for when IA sets aside its rules (*e.g.*, NEL), only one user has access to the scan at any one point in time.

71. IA has developed an initiative known as "Open Libraries" to increase the number of users who can read a given scan at one time. I refer to this number as the "maximum eligible concurrent loan limit" for the scan. In the Open Libraries initiative, IA partners with libraries to leverage their holdings of physical books to, in effect, count their print copies as additional copies that IA may lend as digital books, without doing any re-scanning.

72. At a high level, the process works as follows: (1) A partner library provides a list of ISBNs of its physical book holdings to IA. (2) IA compares these ISBNs to the ISBNs of the scans in the inlibrary collection. This comparison of lists to determine the items they have in common is referred to as an "overlap analysis." (3) For each ISBN that overlaps, IA increases its concurrent loan limit for the associated scan by one.

73. In this way, IA considers a scan to be backed by multiple physical copies: the one that IA photographed and placed in storage, and one or more other different physical books with the same ISBN numbers purportedly maintained in the collections of partner libraries. IA then allows multiple users to have access to copies of a scan concurrently. The maximum eligible concurrent loan limit is equal to one plus the number of partner libraries holding a book with the same ISBN.

74. When the maximum eligible concurrent loan limit is more than one, IA provides the scan for 14-day loans in addition to the 1-hour loans. Each user who takes a loan receives a separate copy of the same scan. Thus, when two users have a loan of a scan at the same time, each is reading a separate copy sent to their device.

75. The Open Libraries program does not involve scanning the physical books held by partner libraries, but instead involves only comparing their ISBNs. The partner library's copy of a book may be in substantially different physical condition than the copy IA scanned.

76. In many cases, IA assigns multiple ISBNs to a single scan, even though only one 10-digit and one 13-digit ISBN correlates to its print edition. These ISBNs assigned by IA may represent different editions of a book. Each additional ISBN that IA assigns to a scan provides an additional chance for a match with a partner library's holdings, thus inflating the matches found in overlap analyses. For example, suppose IA has assigned ISBN A and ISBN B to a scan, with ISBN A being the ISBN appropriate for the physical copy in storage, and ISBN B being the ISBN for a different edition of the same work. If a library's holdings include ISBN B, which is a different edition than IA has in storage (and perhaps a less desirable or older edition), IA nonetheless considers it a match to the scan of ISBN A and increase the scan's concurrent loan limit by one.

77. The design of the IA system for adjusting concurrent loan limits based on overlap analyses does not take into account the fact that a partner library's holdings may be borrowed by the partner's patrons or that old books may be discarded. IA's system is concerned only with whether a book is recorded as being in a partner library's holdings, not with whether that book presently is available for borrowing. In using the Open Libraries initiative to sweep in partner libraries' holdings as extensions of IA's holdings, IA's emphasis is on what the partner library's records indicate the library *owns*, not what the partner library currently has in its stacks for borrowing. To take into account a partner library's lending to its own patrons, IA would need to run overlap analyses in tight synchronization with the library's activities (i.e., stay in real-time synch with which books actually are on the library's shelves). As I will show next, in practice, it appears that the cadence of overlap analyses is inconsistent, as are updates to the library's records. Moreover, there is no reporting from IA to the partner library as users borrow the digital copies from IA—which could lead to a patron borrowing the library's physical copy at the same time one of IA's users borrows the digital copy attributed to that physical copy.

78. IA maintains a collection on its site, named openlibraries, that contains the partner library collections. IA also maintains a collection for each Open Libraries partner library that provides information about activities involving that partner library. Each collection has a unique identifier, such as denverpubliclibrary-ol for the Denver Public Library, and it is at the URL https://archive.org/download/<lib>-ol, where <lib> is the partner library name. IA makes publicly available the results of the overlap analyses that it performs for libraries who participate in the open library program. Curiously, every one of the 17 overlap analyses for "denverpubliclibrary," from 2019/02/13 to 2022/01/01, reports the same number, 100382.

79. I performed a similar analysis for each of the IA partner libraries for which information is available via files of the type just discussed. As set forth below, I find that the majority of them show no change in the number of ISBNs reported.

80. I have analyzed the current reported size of each of these 87 partner library collections. A total of 31 collections are reported to be empty. The University of Arizona Libraries collection (universityofarizona-ol) is the largest, with a reported 379,113 items. Among partners with a non-zero number of items, the median reported collection size is 66743. There are 19 partners with more than reported 100,000 items in their collections.

81. For each partner library, IA makes available to the public a set of files pertaining to the partner's collection. Each summary file indicates the number of ISBNs in the partner library's holdings that were provided to IA for the overlap analysis.

82. For each library in the Open Libraries program (that is, for each library listed in the openlibraries collection), I have used files on the library's Files page, to determine the dates on which overlap analyses were carried out. I also have determined the dates on which the library's

holdings list changed in size, as indicated by the "Total unique ISBNs from <collection_name>" number in the overlap summary file.

83. *Figure 8* below indicates with a dot the date of each overlap analysis for each partner library. A star is used in place of a dot to show dates on which the size of the library's holdings list changed.

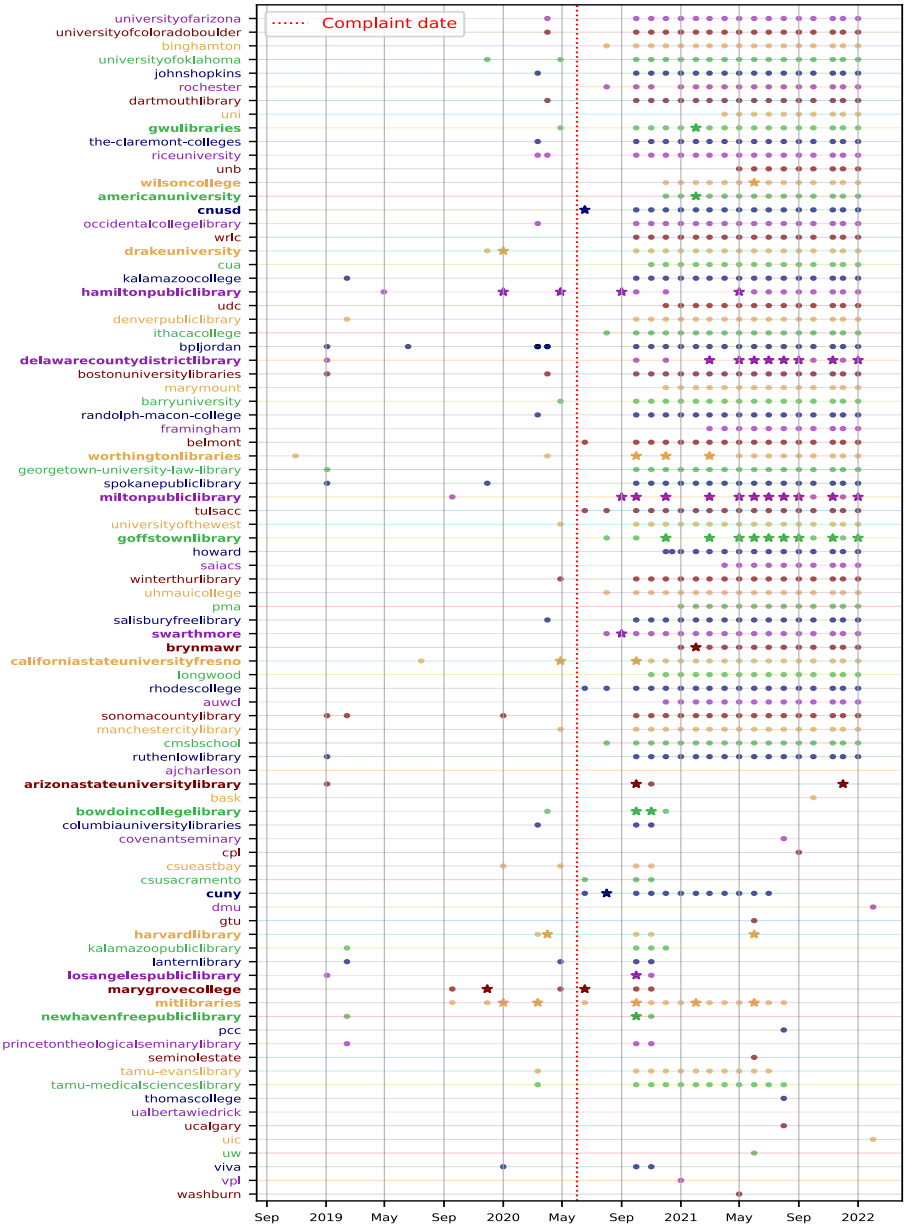


Figure 8: Dots represent overlap analysis reports for partner libraries; stars represent reports that indicate a change in collection size. Those libraries that report at least one change in size are in boldface.

84. From *Figure 8*, it is clear that IA instituted a routine of monthly overlap analyses for most partners in late 2020. The red dotted line indicates the June 1, 2020 date the complaint in this matter was filed. Prior to late 2020, overlap analyses were sporadic and few in number for the current partners. The relative scarcity of stars in the figure reveals that it in general it is uncommon for IA to use updated holdings lists from partners. Indeed, changes in the holding lists are reported for only 21 (24%) of the 87 partners, and only in 60 (6%) of all 992 reports. Only three libraries (Delaware County District Library, Milton Public Library, and Goffs Town Library) show regular and frequent changes in holdings size.

85. After running an overlap analysis with a partner library, IA puts the results of the analysis into effect by updating the "Holdings" list. This is a list that IA maintains and, as described above, consults to determine the concurrent loan limit for a given scan. IA adds the partner to the Holdings list in association with each inlibrary scan identifier for which the partner had a match.

D. Metadata Maintenance

86. "Metadata" generally refers to data concerning other data. Websites offering media items such as music, video, books, and images invariably maintain some metadata about those items. Titles, dates, names of creators, descriptions, comments, ratings, and so on are all examples of metadata encountered by users of services as diverse as Spotify (music), YouTube (video), Amazon (Kindle ebooks), and Instagram (images). Metadata are crucial to the presentation, discovery, and selection of content. When users browse or search services offering media, they commonly are browsing and searching the metadata, not the content that the metadata describe. If an item's metadata are incorrect, users may not be able to find it. Metadata are also crucial to a service's tracking of usage. Imagine a case in which a song is mislabeled with the wrong title and artist on a music service: users will not find it under the expected names, and compensation for

streams of the song may go to the wrong rightsholder. For reasons such as these, quality and accuracy of certain key metadata are of substantial significance to sites offering media.

87. IA's frontend and backend systems related to its Book System rely upon the use of metadata regarding scans (*i.e.*, the digitized version of the print book with a designated scan identifier). These metadata include the fields of information displayed on a details page for each item on the IA Site, such as the name of the book and its author; topics; publisher; publication date; collections(s) into which IA designated the item; and whether it is an access-restricted item. Some of the metadata, such as title, author, and ISBN serve to identify a book; these are identifying metadata. Other metadata, such as the list of collections to which the book belongs, have other purposes. Together, these metadata facilitate a number of functions, such as finding the book by keyword query and dictating if IA's system will provide the book for permanent download rather than only temporary viewing / download.

88. In the cases of scans for which the metadata that IA uses is inaccurate, IA's frontend and backend systems do not function reliably with respect to features that rely upon the inaccurate metadata. When the metadata in question are identifying metadata, such as title, author, and ISBN, one book may be confused with another. This can lead to inaccuracies in other metadata, such as the list of collections to which the book is assigned. For example, inaccurate metadata can lead to situations such as the following:

a. IA delivering a copy of Book A to someone who wants Book B;

b. IA designating Book A for viewing or downloading on an unrestricted basis, rather than designating it into its inlibrary collection for temporary viewing / download;

c. IA exceeding its intended lending limits for Book A by confusing it with Book B;

d. IA under-counting loans (eligible concurrent, actual concurrent, and total) for Book A by attributing those loans to Book B; and

e. IA having an incorrect accounting of its digital holdings and how its digital holdings compare to a library's physical holdings.

89. I have encountered several examples of IA's metadata substantially misidentifying books, which I detail below. IA does not have in place extensive quality-control systems to proactively detect and correct such misidentifications. Consequently, IA's Book System in cases includes and operates based on flawed identification metadata, which leads to problems.

90. Public-facing pages on the IA Site reveal the existence of metadata inaccuracies. An example of public-facing pages of the IA Site revealing the existence of problematic metadata can be seen at https://archive.org/details/thomasberrym00patt. Here, the item is a digitized copy of the book "Total Control" by the author David Baldacci. However, the metadata on the details page is for a different book, entitled "The Thomas Berryman number" by James Patterson: see *Figure 9*.

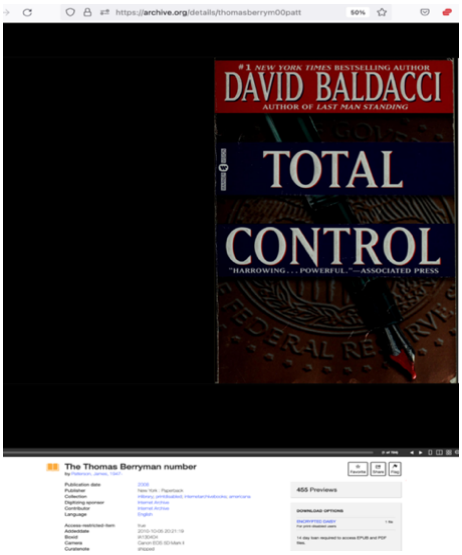


Figure 9: The scan with identifier "thomasberrym00patt" has metadata for "The Thomas Berryman number" but is in fact a scan of Total Control by David Baldacci.

91. By way of another example, *Figure 10* shows a screenshot of a details page on the IA Site for identifier "isbn_9780439799256." The metadata is for the C.S. Lewis book entitled *The Lion, the Witch, and the Wardrobe*, but the scan is a copy of a different book, entitled *Measle and the Wrathmonk*. Because the C.S. Lewis book is a work in suit, IA produced a copy of the scan. I reviewed the scan and it is indeed a copy of *Measle and the Wrathmonk*. As with the James Patterson / David Baldacci example above, this example shows how IA's records can indicate that a scan of Book A is a scan of Book B, and thus deliver books inconsistent with its lending limits and lending counts.

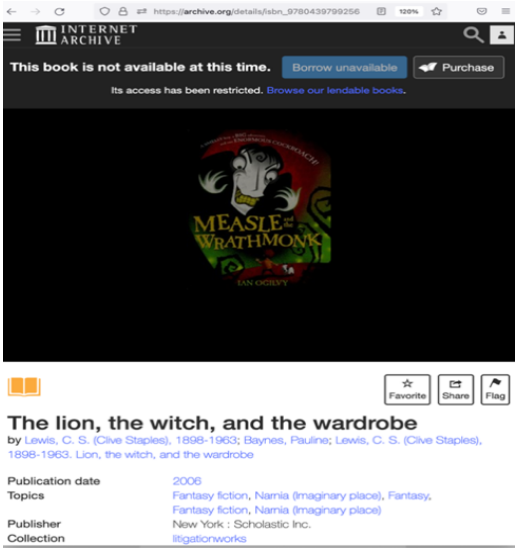


Figure 10: The cover shown on the IA Site for the book with identifier "isbn_9780439799256" does not match its metadata.

92. In addition to the foregoing, IA's metadata has excess ISBNs in the ISBN field, where IA sometimes has multiple different ISBNs for a single digitized copy of a print book.

93. For instance, identifier "psiloveyou00aherx" is for the book *P.S. I Love You* by Cecelia Ahern. The scan is a copy of the paperback version, published in 2005, with the ISBN 0786890754. IA's metadata for that scan lists multiple different ISBNs, including both 0786890754 (the

paperback published in 2005) and 9781401300906. The ISBN 9781401300906 is a hardcover copy, published in 2004.

94. The identifier "isbn_9780316101691" is for the book *Middle School: The Worst Years of My Life* by James Patterson. The scan is a copy of the hardcover version, published in 2011, with the ISBN 9780316101875. IA's metadata for that scan lists multiple different ISBNs, including both 9780316101875 (the hardcover published in 2011) and 9780316101691 (a paperback copy, published in 2012 and the ISBN reflected in the identifier title).

95. Where there are excess ISBNs in the ISBN field, this results in confusion as to which of the ISBNs actually corresponds to the scan. In such situations, IA will have an unclear understanding as to the contents of its digital holdings; *i.e.*, IA will identify its scan as corresponding with any or all of the multiple ISBNs listed in the ISBN field. This is significant because, as discussed earlier, the process of increasing the concurrent lending limit for a scan depends on matching of ISBNs with the holdings of partner libraries.

VIII. Plaintiffs' Works

A. Data and scan productions

96. Plaintiffs attached as Exhibit A to the Complaint in this matter a numbered list of 127 WIS, providing for each a title, author, copyright registration number, and publisher.

97. IA subsequently identified to Plaintiffs hundreds of scans in its Book System responsive to requests for documents pertaining to the 127 WIS. For each of these scans it provided data in files with conventional formats known as CSV and JSON. The CSV files contain records of updates made to IA's files and metadata regarding a scan. IA refers to these records as the "catalog task history" for the scan. The JSON file provides a wide range of additional metadata about the scan.

98. I matched the 127 WIS to the scans that IA produced in discovery from its Book System. The results of this matching process are provided in the Excel file named "Tabulation.xlsx" that I prepared with my reports. This is a comprehensive table that presents my association of scans with the WIS and various data derived from IA's JSON and CSV files.

99. The Tabulation table contains 376 data rows reflecting the association of 371 distinct scans with 127 WIS. Due to its large number of columns, the Tabulation table is not well suited for printing or presentation in a document. Instead, I have attached to this declaration various exhibits that are based upon the Tabulation table, as described here and in the next section.

100. The first exhibit I present based on the Tabulation table is simply a list associating the WIS with scans. **Exhibit 1** (Exhibit T2 in my reports) is an association of the WIS to the scans provided by IA. For each WIS, the first three columns are from Exhibit A to the Complaint, and the next column lists the one or more matching scan identifier(s). From this, one can see that IA has one or two scans for some WIS and up to over 10 or 20 scans for others. The last two columns provide, in addition, information about the Contributor and Digitizing Sponsor. The Contributor is the person or entity that provided (and owns) the physical book being digitized. The Digitizing Sponsor is the person or entity who paid for the digitization.

B. History of Internet Archive's Usage of Works-in-Suit

101. Based on the Tabulation table, I have prepared several exhibits reflecting IA's usage of the WIS. However, I note that the loan data from the CSV files are incomplete. According to the February 14, 2022 declaration of Brenton Cheng, a Senior Engineer at IA, these contain loan information from on or around March 2017 though on or around September 2, 2020 and exclude a few short durations ("generally less than a couple days") within this period. Also, I note that IA ceased lending the WIS scans in the period following Plaintiffs' filing the complaint in this matter

on June 1, 2020. Thus, the loan data that IA provided excludes data on any loans prior to March 2017 or after around early September 2020.

102. My exhibits reflecting IA's usage of WIS include the following:

103. **Exhibit 2** (Exhibit T3 in my reports) sets forth the maximum actual concurrent loans and the total loans for each Scan that matches with a WIS, from March 2017 to on or about Sept 2, 2020 (the earliest and latest loan dates, respectively, for which IA provided data). Data are divided into three periods: prior to 2019, 2019, and 2020. There is also a column that provides the overall number of loans. In Exhibit 2, the maximum actual concurrent loans data shows many instances in which IA provided a scan to more than one user at a time. By way of example, as to the book *Ship Breaker* by Paolo Bacigalupi (WIS # 4), while IA has only a single scan, IA provided 5 users with copies at a time in 2019. Moreover, IA provided up to 15 users at a time with copies in 2020.

104. **Exhibit 3** (Exhibit T5 in my reports) sets forth total loans for each WIS, during the period from March 2017 to on or about Sept 2, 2020 (the earliest and latest loan dates, respectively, for which IA provided data). As with the other exhibits, the loan data that IA provided does not include loans that occurred prior to March 2017 or after early September 2020. Numbers are obtained by summing across all scans that match that WIS, for all available data, with the number of such scans also specified. The lending totals vary by WIS, with some having 1,000 or more. The total number of loans of WIS is 46,307.

105. **Exhibit 4** (Exhibit T7 in my reports) sets forth the number of de-duplicated ISBNs and additions to maximum eligible concurrent loans for each scan that matches with a WIS. As with the other exhibits, the loan data that IA provided does not include loans that occurred prior to March 2017 or after early September 2020. Exhibit 4 is thus based on loan data only for the period roughly from March 2017 to early September 2020. In Exhibit 4, the additions to the maximum

eligible concurrent loans are due to data matches with collections of partner libraries in IA's Open Libraries initiative. The more ISBNs that IA associates with a single scan, the greater the opportunity that the scan has to match with ISBNs in partner collections. This exhibit shows how IA scales a single scan of a particular physical book in its possession into a much larger dissemination scheme that involves copies of that scan being provided to multiple users at a time. In an abundance of caution, following questions in my deposition that certain apparent partner library collections do not count toward increasing the maximum number of concurrent loans, I have prepared an alternative version of Exhibit 4, labelled as **Exhibit 4A** (an alternative version of Exhibit T7 in my reports), to exclude them. Specifically, this version excludes partner libraries lacking an "-ol" at the end of their collection name, as well as the collection named "stmaryscountylibrary-ol."

106. **Exhibit 5** (Exhibit T8 in my reports) sets forth Number of de-duplicated ISBNs and additions to maximum eligible concurrent loans for each WIS—obtained by summing across all Scans that match that WIS. Exhibit 5 is likewise based on loan data only for the period roughly from March 2017 to early September 2020. Exhibit 5 is a WIS-level view of Exhibit 4, summing over the scans comprising each WIS. This again shows the effect on availability of a work simply by having more scans of the work. Just as with Exhibit 4, I have prepared an alternative version of Exhibit 5, labelled as Exhibit 5A (an alternative version of Exhibit T8 in my reports), that excludes certain partner library collections.

107. **Exhibit 6** (Exhibit T9 in my reports) sets forth maximum actual concurrent loans, total loans, scan date, and inlibrary date, for each scan associated with a WIS, during the period from March 2017 to on or about Sept 2, 2020 (the earliest and latest loan dates, respectively, for which IA provided data). Exhibit 6 reveals the dates when the scans associated with the WIS were

scanned and put in the inlibrary collection for lending. All of the WIS were included in the inlibrary (i.e., "Books to Borrow") collection and the subject of loans. The two highest quarters of IA's scanning of WIS were Q4 2019 and Q1 2020.

C. Additional Data Regarding Usage Patterns—General and of WIS

108. In the final few weeks before my report was due, IA provided two declarations dated January 31, 2022 and February 14, 2022 by Brenton Cheng (who, as mentioned above, is a Senior Engineer at IA) and produced associated data and source code. These materials contain information regarding, among other topics, usage activity of scans in general and usage activity of scans containing the WIS in particular.

Download and Loan Counts Across the Contents of the inlibrary Collection

109. I have used the data to confirm how often loans involve downloads to EPUB or PDF. In making that calculation, I do not divide downloads by a figure that is the sum of 14-day and 1-hour loans. Instead, because PDF and EPUB downloads are available only for 14-day loans, a more meaningful is obtained by dividing the number of download attempts by the number of loans that permit such downloads; that is, the number of 14-day loans. In the table below, I present counts of each sort of event across the full date range for which IA produced data along with percentages with respect to 14-day loans. I find that the number of PDF and EPUB download attempts together is 62.8% of the number of 14-day loans. Separately, PDFs are 51.9% and EPUBs are 10.9%.

14-day loans	13,793,385
1-hour loans	31,437,549
PDF download attempts	7,154,753
EPUB download attempts	1,506,489

PDF as % of 14-day loans	51.9%
EPUB as % of 14-day loans	10.9%
PDF and EPUB as % of 14-day loans	62.8%

Loan Analytics Database Output for Works-in-Suit

110. I have reviewed an Excel file that IA produced to Plaintiffs on January 31, 2022, and which, according to Mr. Cheng's description, contains output from a "Loan Analytics Database" reflecting events related to loans from April 25, 2020 onward of scans that IA associated with the WIS. In assessing this data, it is important to note that the loan data would pertain to loan events for only a limited period after April 25, 2020 because IA attempted to remove the WIS from its inlibrary collection shortly after Plaintiffs commenced this lawsuit on June 1, 2020.

111. Even for that limited period, the "Loan Analytics Database" contains data for at least one scan identifier for every WIS. The data indicates at least one download event ("access_pdf" or "access_epub") for every WIS except WIS #72 (*Legend* by Marie Lu) and WIS #73 (*Station Eleven*, by Emily St. John Mandel). These two WIS are unusual and may represent a breakdown in IA's recordkeeping or production of data as to these two WIS, in that the only events in the Loan Analytics Database for their scan identifiers are "expire" events dated May 22, 2020. These expire events consist of 34 for WIS #72 identifier legend00luma, and 212 total for WIS #73 identifiers stationeleven0000mand and stationelevennov0000mand. Based on the other data (62.8% of 14-day loans have a PDF or ePub download), the likelihood of 246 loans but with no engagement or downloading is extremely slim.

D. The WIS are members of a much larger population of Plaintiffs' works that IA scanned and provides to users

112. The WIS consist of 127 works drawn from the catalogs of the four Plaintiffs. I was asked to assess how IA has copied and used books from Plaintiffs' broader "in-print" catalogs, which collectively consist of hundreds of thousands of items. To address this question, I studied the extent to which IA's inlibrary ("Books to Borrow") collection contains scans for works that the four Plaintiffs have in their catalogs in both print (hardcover or paperback) and electronic forms. For this project, I made use of catalogs for Hachette Book Group (67,550 items, HACHETTE0002576), HarperCollins Publishers (112,221 items, HC0003508), Penguin Random House (190,471 items, PRH0068115), and John Wiley & Sons (157,110 items, WILEY0011680). I understand that these in-print catalogs reflect items that the Plaintiffs currently make commercially available. Each item in the catalog has a distinct ISBN and represents the publication of a work in particular format. For example, a work (that is, a given title by a given author) available in one hardcover, one paperback, and one electronic format would have a distinct ISBN for each format and appear three times in the catalog.

113. To accomplish this analysis, I wrote and executed several programs to process Plaintiffs' catalogs, determine the availability of Plaintiffs' works for borrowing on the IA Site, and present the results. I first used a program that reads a publisher's catalog file and outputs, for each work for which the catalog contains entries for both physical (hardcover or paperback) and electronic copies, a line containing a type, either P (physical) or E (electronic) plus the item's ISBN, author, title, and publication date. I then used a second program that queries the IA Site to determine whether each of the physical and electronic ISBNs for the qualifying works is present in IA's Books to Borrow collection. I then used a third program to take these results and produce a report on the results of the lookups.

114. Below, I present the results for two publication date ranges: all available dates, and all dates through 2015. I chose 2015 because, as I described in an earlier section, IA's policy calls for it not to add to the inlibrary collection scans of books published in the past five years. Thus, in searching the inlibrary collection for works in the Plaintiffs' catalogs, it is to be expected that works published in 2016 or later will be absent.

115. Based on my analysis, I found that the following percentages of Plaintiffs' qualifying works were present in the IA's inlibrary "Books to Borrow" collection across all publication dates:

f. 34.6% for Hachette Book Group
g. 36.7% for HarperCollins
h. 35.8% for Penguin Random House
i. 15.5% for John Wiley & Sons

116. Based on my analysis, I found that the following percentages of Plaintiffs' qualifying works (titles in their catalogs in both print and electronic forms) were present in the IA Site's inlibrary "Books to Borrow" collection for all publication dates through 2015:

j. 54.6 for Hachette Book Group
k. 48.4% for Penguin Random House
l. 21.3% for John Wiley & Sons

(I provide a percentage for HarperCollins titles only across all publication dates, as I did not have the publication dates for HarperCollins titles.)

117. In Table 1 below, I present these results with the underlying numbers on which they are based.

Harper	7055	19205	36.7	n/a	n/a	n/a
Penguin	16496	46059	35.8	15746	32520	48.4
Wiley	4933	31870	15.5	4889	22985	21.3

118. As a supplement to the above analysis, I examined the dates on which IA cre found, per the preceding analysis, to be present the inlibrary collection and conta qualifying works from their catalogs. I used a program to obtain from the IA Site record (in JSON format) for each of the scans. I then used a program to extract the s plot their distribution for each publisher. Each of the figures below (*Figure 11 to Figu* the number of scans created by IA each month between mid 2009 and the present (Fe for the qualifying works in the indicated publisher's catalog. These figures reveal that of Plaintiffs' works to the inlibrary collection has increased substantially since the date of the Complaint in this matter.

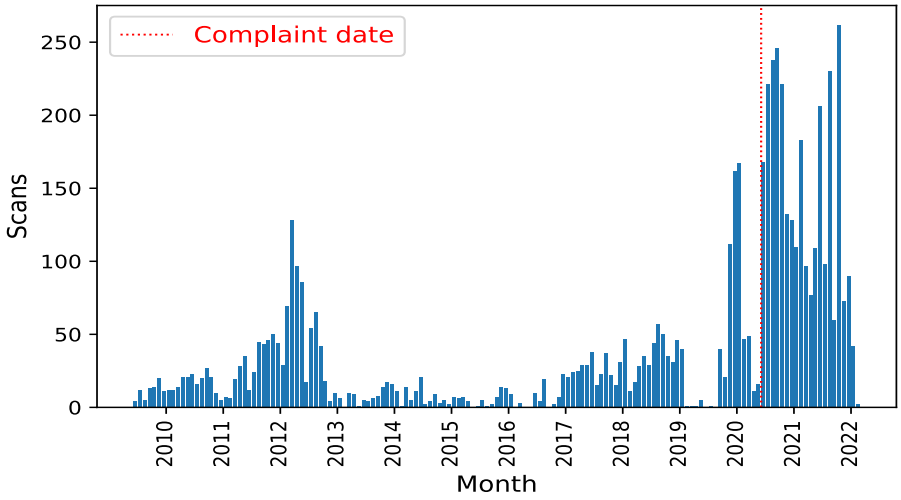


Figure 11: Distribution indicating the number of inlibrary scans made by IA each month of qualifyin Hachette catalog. Post-Complaint: 2,993 total, with an average of 143/month over 21 months.

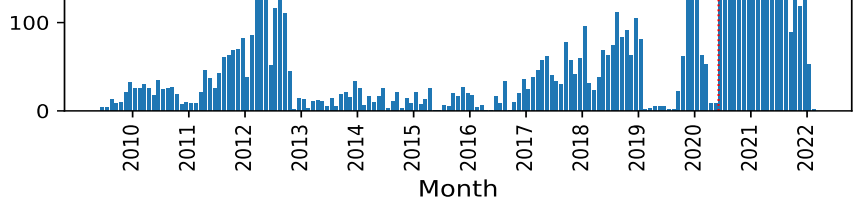


Figure 12: Distribution indicating the number of inlibrary scans made by IA each month of qualifyin Harper catalog. Post-Complaint: 4,171 total, with an average of 199/month over 21 months

.

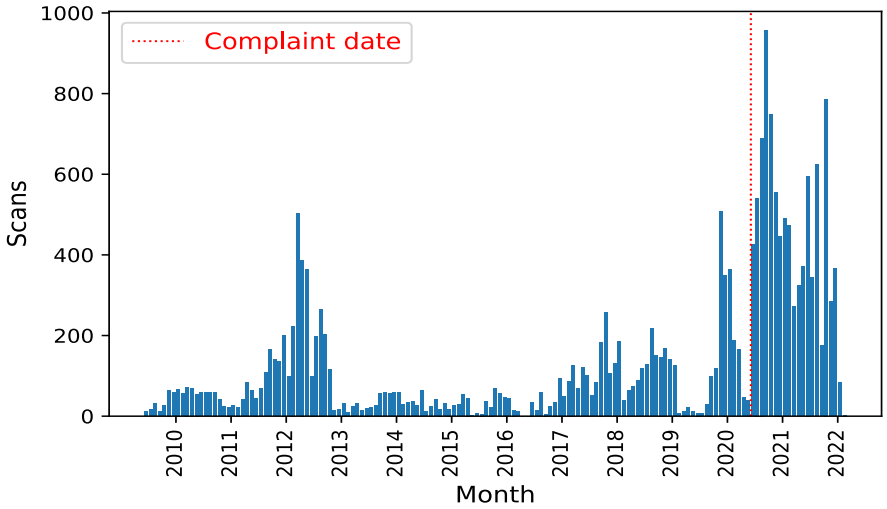


Figure 13: Distribution indicating the number of inlibrary scans made by IA each month of qualifyin Penguin catalog. Post-Complaint: 9,558 total, with an average of 455/month over 21 months.

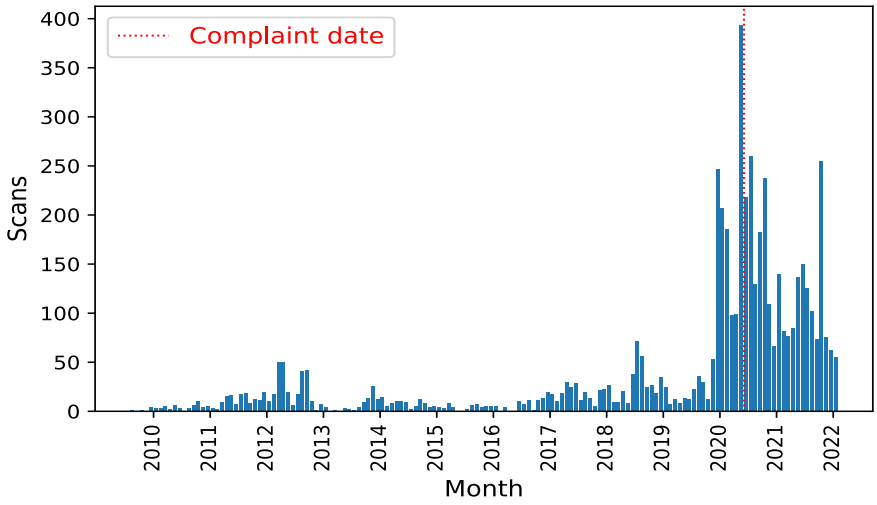


Figure 14: Distribution indicating the number of inlibrary scans made by IA each month of qualifyin Wiley catalog. Post-Complaint: 2,622 total, with an average of 125/month over 21 months.

IX. IA's Response to Takedown Notices and the Complaint

119. Plaintiffs or their representatives have sent copyright notices to IA identifying unauthorized copies of Plaintiffs' works to be removed. In my work on this matter, I reviewed IA's technical capabilities to receive and respond to such copyright notices, also known as takedown notices, and how IA's actions in response to notices compare to its capabilities.

120. When IA removes a scan from its in-library collection in response to a takedown notice, from a technical matter, it is done by employees with certain permissions (also referred to as "administrative privileges"). In taking action to remove or disable access to an alleged infringement, IA's employee accesses a web page associated with the scan named in the complaint called its "manage page." This page has, for a scan with identifier <id>, the address "https://archive.org/manage/<id>". A scan's manage page allows a person with administrative privileges to check a box or click a button to change various aspects of the scan's status. This is done using an interface that IA refers to as the "book dashboard."

121. The administrative interface that IA calls "ROMM"—standing for Revenge of MetaManager—which also is known as the book dashboard, allows for addressing a single identifier or multiple scan identifiers and applying tasks. That tool also allows the user to search by publisher name, author name or other fields and then to apply tasks against the results. In past years, an IA employee would need to edit the scan's metadata via an "edit XML page", with XML being the format in which the metadata are maintained. But it is no longer necessary to manually edit metadata fields on the XML page, because an IA employee can now use the book dashboard to effect the change. The book dashboard is accessed via a scan's manage page.

122. I have observed that, as to copyright notices IA received before the lawsuit, IA failed to remove or disable access to certain of the alleged infringements cited in those notices, both for WIS and for other works. I have observed this based on a review of certain takedown notices

produced in this case, my use of the IA Site, and the work of Tracy Offner, as described in her Declaration dated February 7, 2022 ("Offner Decl.").

123. One of the WIS is the book *Franny & Zooey* by J.D. Salinger. The identifier "frannyzooey00saliric" corresponds to that book, and it was listed (including by identifier) in copyright notices sent to IA on July 5, 2018 and November 9, 2018. Even though it appeared in those two prior notices, IA did not remove or disable access to its scanned copy. Just before commencing this lawsuit, IA's scanned copy of *Franny & Zooey* by J.D. Salinger, with the same identifier, "frannyzooey00saliric," remained available for download, and was indeed downloaded, from IA's "inlibrary" collection. *See* Offner Decl. ¶ 118.

124. Another WIS is *The Mysterious Benedict Society* by Trenton Lee Stewart. The identifier "mysteriousbenedi00stew" corresponds to that book, and it was listed (including by identifier) in 35 copyright notices sent to IA between March 26, 2018 and May 22, 2018. Nonetheless, IA did not remove or disable access to its scanned copy. Just before Plaintiffs commenced this lawsuit, IA's scanned copy of *The Mysterious Benedict Society* by Trenton Lee Stewart, with the identifier "mysteriousbenedi00stew," was available for download, and was indeed downloaded, from IA's "inlibrary" collection. *See* Offner Decl. ¶ 129.

125. One of the WIS is *The Mysterious Benedict Society and the Perilous Journey* by Trenton Lee Stewart. The identifier "mysteriousbenedi00stew_0" corresponds to the book, and it was listed (including by identifier) in 36 copyright notices between March 26, 2018 and May 22, 2020. Nonetheless, IA did not remove or disable access to its scanned copy. Just before Plaintiffs commenced this lawsuit, IA's scanned copy of *The Mysterious Benedict Society and the Perilous Journey* by Trenton Lee Stewart, with the identifier "mysteriousbenedi00stew_0," remained

available for download, and was indeed downloaded, from IA's "inlibrary" collection. *See* Offner Decl. ¶ 131.

126. WIS also include the following books written by author James Patterson: (a) *I Funny: A Middle School Story*; (b) *Invisible*; and (c) *Middle School: The Worst Years of My Life*. On April 13, 2020, Hachette emailed IA requesting that it remove or disable access to all James Patterson books. IA responded on April 14, 2020 acknowledging the email and confirming a good faith effort to identify and disable lending access for James Patterson books. Nonetheless, the following WIS remained available for download, and were indeed downloaded, from Defendant's "inlibrary" collection on May 20-21, 2020:

i. *I Funny: A Middle School Story* by James Patterson, with the identifier "ifunny0000patt_w7e5."

ii. *I Funny: A Middle School Story* by James Patterson, with the identifier "ifunny0000patt."

iii. *Invisible* by James Patterson and David Ellis, with the identifier isbn_9780316405348."

iv. *Middle School: The Worst Years of My Life* by James Patterson, with the identifier "isbn_9780316101691."

v. *Middle School: The Worst Years of My Life* by James Patterson, with the identifier "middleschoolwors0000patt."

See Offner Decl. ¶ 126.

127. Besides works specifically listed in the Complaint, IA failed to remove or disable access to other works that were cited, including by identifier, in numerous copyright notices. Below are some examples, involving authors who have at least one title specifically listed in the Complaint.

128. The URL http://archive.org/details/treasureworthsee00brow corresponds to the book *A Treasure Worth Seeking* by Sandra Brown. The identifier "treasureworthsee00brow" was subject

to 79 take down notices from January 23, 2018 to May 22, 2018. I confirmed that, as of February 22, 2022, the book was available to borrow via the "inlibrary" collection.

129. The URL https://archive.org/details/temptationskiss00sand corresponds to the book *Temptation's Kiss* by Sandra Brown. The identifier "temptationskiss00sand" was subject to 88 take down notices from December 29, 2017 to May 22, 2018. I confirmed that, as of February 22, 2022, the book was available to borrow via the "inlibrary" collection.

130. The URL http://archive.org/details/slowheatinheaven00sand corresponds to the book *Slow Heat in Heaven* by Sandra Brown. The identifier "slowheatinheaven00sand" was subject to 92 take down notices from December 15, 2017 to May 22, 2018. I confirmed that, as of February 22, 2022, the book was available to borrow via the "inlibrary" collection

131. The URL http://archive.org/details/rosiedunne00aher corresponds to the book *Rosie Dunne* by Cecelia Ahern. The identifier "rosiedunne00aher" was subject to 75 take down notices from January 29, 2018 to May 22, 2018. I confirmed that, as of February 22, 2022, the book was available to borrow via the "inlibrary" collection.

132. By way of further example, on April 6, 2020, Hachette sent IA a request to remove all works for author Jonathan Safran Foer. Table 2 lists those that appear in the inlibrary collection as of February 19, 2022:

Table 2: Copies of books by Jonathan Safran Foer in the IA inlibrary collection as of February 19, 2022.

	Title	**Author**
https://archive.org/details/everythingisillu0000foer	Everything is Illuminated	Jonathan Safran Foer
https://archive.org/details/hereiam0000foer_h8m6	Here I Am	Jonathan Safran Foer
https://archive.org/details/extremelyloudinc0000foer_h8a2	Extremely Loud & Incredibly Close	Jonathan Safran Foer
https://archive.org/details/everythingisillu0000foer_k5a0	Everything is Illuminated	Jonathan Safran Foer
https://archive.org/details/extremelyloudinc0000foer_u8u5	Extremely Loud & Incredibly Close	Jonathan Safran Foer
https://archive.org/details/everythingisillu0000foer_p5l2	Everything is Illuminated	Jonathan Safran Foer
https://archive.org/details/hereiam0000foer_t0u7	Here I Am	Jonathan Safran Foer
https://archive.org/details/hereiam0000foer_u9e5	Here I Am	Jonathan Safran Foer

133. On July 23, 2014, Plaintiff Penguin Random House ("PRH") emailed its catalog to IA in an Excel spreadsheet, demanding that IA remove "[u]nauthorized, scanned versions of Penguin Random House titles being lent to patrons... immediately." The email added that the "attached file includes both print and eBook isbn for all Random House titles, and print isbns (only) for Penguin titles. It includes close to 60,000 unique isbns. Please let us know the time frame in which we can expect these titles to be removed, and if you need anything additional from us to facilitate this process."

134. I assessed whether scanned copies of the books listed in PRH's catalog, as provided to IA in July 2014, were in IA's inlibrary collection at the time of my analysis in 2022. I did this by querying the IA Site for availability in the inlibrary collection of the books listed in that spreadsheet. Of the 58,754 ISBNs listed in the July 2014 PRH catalog, 35,413 (60.3%) remain in PRH's current commercial catalog, and 23,341 (39.7%) do not. Looking at what fraction of these books are to be found in IA's inlibrary collection, I determined that the inlibrary collection contained:

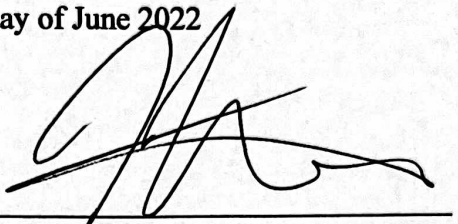
- 47.1% (16,681[1] of 35,413) of the 2014 catalog books that are in the current catalog;
- 48.7% (11,366 of 23,341) of the 2014 catalog books that are not in the current catalog; and
- 47.7% (28,047 of 58,754) of all 2014 catalog books.

135. Based on the observations above, I conclude that IA's actions with respect to copyright notices have been fallen short of its technical capabilities.

[1] I note that 12,318 of these 16,681 works are "qualifying works" in the sense used in this report, i.e., works for which there is both a physical and electronic ISBN in the PRH catalog.

I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct.

Executed in Chicago, Illinois this 29th day of June 2022



Dr. Ian Foster